Aligning AI With Shared Human Values

Year: 2020 | Citations: 723 | Authors: Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, J. Li

---

Abstract

We show how to assess a language model's knowledge of basic concepts of morality. We introduce the ETHICS dataset, a new benchmark that spans concepts in justice, well-being, duties, virtues, and commonsense morality. Models predict widespread moral judgments about diverse text scenarios. This requires connecting physical and social world knowledge to value judgements, a capability that may enable us to steer chatbot outputs or eventually regularize open-ended reinforcement learning agents. With the ETHICS dataset, we find that current language models have a promising but incomplete understanding of basic ethical knowledge. Our work shows that progress can be made on machine ethics today, and it provides a steppingstone toward AI that is aligned with human values.